Regularised Energy Model for Robust Monocular Ego-motion Estimation

Hsiang-Jen Chien and Reinhard Klette
School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology, Auckland, New Zealand

Keywords: Visual Odometry, Camera Motion Recovery, Perspective-n-points Problem, Nonlinear Energy Minimisation.

Abstract: For two decades, ego-motion estimation is an actively developing topic in computer vision and robotics. The principle of existing motion estimation techniques relies on the minimisation of an energy function based on re-projection errors. In this paper we augment such an energy function by introducing an epipolar-geometry-derived regularisation term. The experiments prove that, by taking soft constraints into account, a more reliable motion estimation is achieved. It also shows that the implementation presented in this paper is able to achieve a remarkable accuracy comparative to the stereo vision approaches, with an overall drift maintained under 2% over hundreds of metres.

1 INTRODUCTION

Recovering camera motion from imagery data is one of the fundamental problems in computer vision. Image-based motion estimation provides a complementary solution to GPS-engaged positioning systems which might fail in close-range (e.g. indoor) environments or due to any circumstances without clear satellite signals. A variety of techniques can be found in a number of applications in the context of *simultaneously localisation and mapping* (SLAM) (Konolige et al., 2008), *structure from motion* (SfM), or *visual odometry* (VO) (Scaramuzza and Fraundorfer, 2011).

The estimation of camera motion can be achieved in different ways, up to the availability of inter-frame point correspondences. In the case of ToF or RGB-D cameras where the pixel depths are available, the relative pose of the sensor between two different frames can be derived using 3D-to-3D correspondences by means of rigid body registration (Hu et al., 2012). It is a more general case where the 3D coordinates of pixels are known only in the previous frame with their locations observed in the current frame. In such a case the ego-motion is estimated from 3D-to-2D correspondences, and the minimisation of the deviations of the projected 3D coordinates from the observed 2D locations has been proven to be the golden standard solution to the ego-motion estimation problem (Engels et al., 2006).

In this paper we provide a quick review for the underlying mathematical models of the monocular ego-motion estimation problem. Based on these models, we propose an augmented energy function that regularises the iterative adjustment of estimated ego-motion by taking epipolar constraints into account.

The paper is organised as follows. In Section 2 we provide a literature review on mathematical foundations of the monocular ego-motion estimation problem. In Section 3 a revised energy model is proposed which is then verified by the experiments reported in Section 4. We conclude this paper in Section 5.

2 MONOCULAR EGO-MOTION

We review the common process by starting with the theory, and ending with comments on implementation.

Theory. Following the pinhole camera projection model, a 3D point $P = (x, y, z)$ is projected into a pixel location (u, v) in the image plane by

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} \sim \begin{pmatrix} f_u & 0 & u_c & 0 \\ 0 & f_v & v_c & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} = (\mathbf{K}\,\mathbf{0}) \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \quad (1)$$

where the upper 3×3 triangular matrix $\mathbf{K}$ is the *camera matrix* modelled by the intrinsic parameters of the

Chien H. and Klette R.
Regularised Energy Model for Robust Monocular Ego-motion Estimation.
DOI: 10.5220/0006100303610368
In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)*, pages 361-368
ISBN: 978-989-758-227-1

camera including focal lengths f_u and f_v, and the image centre or principle point (u_c, v_c). By $\sim$ we denote projective equality (i.e. equality up to a scale).

As the camera moves to a new position, the same point P, if it remains stationary, is observed at a different pixel location. The movement of the camera introduces a new coordinate system which can be modelled by an Euclidean transformation with respect to the previous frame. Let $(\mathbf{R}\,\mathbf{t})$ be such a transformation, where $\mathbf{R} \in SO(3)$ is the rotation matrix, and $\mathbf{t} \in \mathbb{R}^3$ is the translation vector. The new projection of point P is found by

$$\begin{pmatrix} u' \\ v' \\ 1 \end{pmatrix} \sim \mathbf{K}\begin{pmatrix} \mathbf{R} & \mathbf{t} \end{pmatrix}\begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \tag{2}$$

3D-to-2D ego-motion estimation algorithms rely on the principle that, given sufficiently many observations $(x,y,z) \leftrightarrow (u',v')$, it is possible to determine the unknown transformation $(\mathbf{R}\,\mathbf{t})$. The estimation of such transformations is known as the *perspective-from-n-points* (PnP) problem (Lepetit et al., 2009).

A linear approach treats the projection as a general linear transform controlled by the 3-by-4 projection matrix $\mathbf{P} = \mathbf{K}\begin{pmatrix} \mathbf{R} & \mathbf{t} \end{pmatrix}$. For each observation $(x,y,z) \leftrightarrow (u',v')$, two linear constraints are obtained as follows:

$$\begin{pmatrix} \mathbf{A} & & \\ & -u'x & -u'y & -u'z \\ & -v'x & -v'y & -v'z \end{pmatrix}\begin{pmatrix} \mathbf{P}_1^\top \\ \mathbf{P}_2^\top \\ \mathbf{P}_3^\top \end{pmatrix} = \mathbf{1} \tag{3}$$

where $\mathbf{P}_i$ denotes the i-th row of the projection matrix $\mathbf{P}$, and

$$\mathbf{A} = \begin{pmatrix} x & y & z & 0 & 0 \\ 0 & 0 & 0 & x & y \end{pmatrix} \tag{4}$$

Having six world-image correspondences, a linear system of twelve unknowns can be constructed. If the observations are linearly independent then the matrix $\mathbf{P}$ can be calculated, allowing one in turn to use the calibrated camera matrix $\mathbf{K}$ to recover the motion by

$$\begin{pmatrix} \mathbf{R} & \mathbf{t} \end{pmatrix} = \mathbf{K}^{-\top}\mathbf{P} \tag{5}$$

In practice more than six correspondences are used to construct an over-determined linear system, and a least-squares-solution yields a more robust solution. This strategy is known as the *direct linear transform* (DLT) method.

As an Euclidean transformation essentially has six *degrees of freedom* (DoF) while there are twelve unknowns in $\mathbf{P}$, the recovered rotation matrix $\mathbf{R}$ is not guaranteed to be a valid element in $SO(3)$ due to over-parameterization. Furthermore, the minimised algebraic errors, subject to Eq. (3), lack of geometric interpretation. To address these issues, a nonlinear adjustment strategy is usually carried out following the linear estimation step.

Assuming that the 3D measurement noise follows a Gaussian model, the *maximum-likelihood estimation* (MLE) of $(\mathbf{R}\,\mathbf{t})$ is achieved by a minimisation of the sum-of-squares of the reprojection error:

$$\phi_{\mathrm{R}}(\mathbf{R},\mathbf{t}) = \sum_i \left\| (u_i', v_i')^\top - \pi_{\mathbf{K}}[\mathbf{R}(x_i, y_i, z_i)^\top + \mathbf{t}] \right\|_{\Sigma_i}^2 \tag{6}$$

where $\pi_{\mathbf{K}} : \mathbb{R}^3 \to \mathbb{R}^2$ is the projection function that maps a 3D point into the projective space P^2 using the camera matrix $\mathbf{K}$. It also converts the resulting homogeneous coordinates into a Cartesian plane. By Σ_i we denote the 2×2 error covariance matrix of the i-th-point correspondence.

As Eq. (6) cannot be solved in any closed form, one may adopt a nonlinear least-squares minimiser, say the Levenberg-Marquardt algorithm (Levenberg, 1944), to minimise the energy function, starting with the solution found using the DLT estimation as an initial guess.

Motion without 3D Prior. For a monocular vision system, 3D coordinates (x,y,z) might not be available as a prior. In this case, the motion of the camera can still be recovered from epipolar conditions but where the scale of $\mathbf{t}$ remains undetermined. Without loss of generality, we assume that $\|\mathbf{t}\| = 1$ in the following context. Let $(u,v) \leftrightarrow (u',v')$ be a 2D-to-2D correspondence. It follows that

$$\begin{pmatrix} u' \\ v' \\ 1 \end{pmatrix}^\top \mathbf{K}^{-\top}[\mathbf{t}]_\times \mathbf{R}\mathbf{K}^{-1}\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = 0 \tag{7}$$

where

$$[\mathbf{t}]_\times = \begin{pmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{pmatrix} \tag{8}$$

denotes the skew-symmetric form of $\mathbf{t} = (t_x, t_y, t_z)^\top$. Equation (7) is the well-known epipolar condition, and the matrix $\mathbf{E} = [\mathbf{t}]_\times \mathbf{R}$ is called the *essential matrix*.

Among a variety of essential matrix recovery techniques, the eight-point algorithm is a popular choice. The method first estimates the *fundamental matrix* $\mathbf{F} = \mathbf{K}^{-\top}\mathbf{E}\mathbf{K}^{-1}$ using at least eight point correspondences. For each correspondence $(u,v) \leftrightarrow (u',v')$, a homogeneous constraint is introduced by Eq. (7) as follows:

$$\begin{aligned} &uu'f_{11} + vu'f_{12} + u'f_{13} + uv'f_{21} + vv'f_{22} \\ &+ v'f_{23} + uf_{31} + vf_{32} + f_{33} = 0 \end{aligned} \tag{9}$$

where f_{ij} denotes an element of the fundamental matrix. By means of linear algebra techniques, all the nine elements of the fundamental matrix can be determined up to a scale using at least eight constraints.

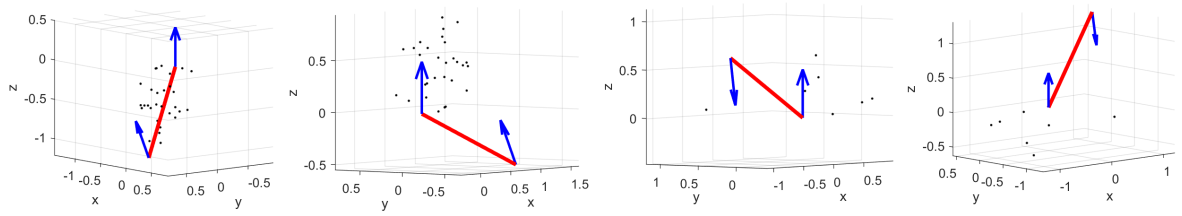

Figure 1: An example of four possible ego-motion estimates from essential matrix decomposition. Only the second to the leftmost solution shows a valid geometric configuration, where all the triangulated 3D points lie in front of both cameras.

According to $\mathbf{E} = \mathbf{K}^\top \mathbf{F} \mathbf{K}$, one may obtain the essential matrix from the solved fundamental matrix.

The motion, denoted by $\mathbf{R}$ and $\mathbf{t}$, can be extracted from a calculated essential matrix $\mathbf{E}$. One may compute a *singular value decomposition* (SVD)

$$\mathbf{E} = \mathbf{U}\mathbf{D}\mathbf{V}^\top \tag{10}$$

of matrix $\mathbf{E}$ where $\mathbf{U}$ and $\mathbf{V}$ are 3×3 orthonormal matrices, and $\mathbf{D} = diag(1,1,0)$ is a diagonal matrix having a 1 as the first and second diagonal element, and 0 as the third (due to the rank deficiency of $\mathbf{E}$). By introducing two matrices

$$\mathbf{Z} = \begin{bmatrix} 0 & \pm 1 & 0 \\ \mp 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} 0 & \mp 1 & 0 \\ \pm 1 & 0 & 0 \\ 0 & 0 & \pm 1 \end{bmatrix} \tag{11}$$

and based on $\mathbf{D} = \mathbf{Z}\mathbf{W}$ and $\mathbf{U}^\top \mathbf{U} = \mathbf{I}$, one may now rewrite Eq. (10) as follows:

$$\mathbf{E} = \mathbf{U}\mathbf{Z}\mathbf{U}^\top \mathbf{U}\mathbf{W}\mathbf{V}^\top \tag{12}$$

It is verified that $\mathbf{S} = \mathbf{U}\mathbf{Z}\mathbf{U}^\top$ is a skew-symmetric matrix, and $\mathbf{R}' = \mathbf{U}\mathbf{W}\mathbf{V}^\top$ is an orthonormal matrix. Following the definition $\mathbf{E} = [\mathbf{t}]_\times \mathbf{R} = \mathbf{S}\mathbf{R}'$, the rotation matrix $\mathbf{R}$ and the unit translation vector $\mathbf{t}$ are instantly found.

Due to the sign ambiguity of $\mathbf{Z}$ and $\mathbf{W}$, there are four possible solutions. As described in the next section, the best candidate is decided by applying a triangulation method on $(u,v) \leftrightarrow (u',v')$, and checking the number of the resulting points that fall *in front of* the cameras to select the best candidate. In the non-singular case, only one candidate gives a valid geometric setup. Figure 1 depicts an example of all the four possible solutions.

Triangulation. Triangulation is the process of computing 3D coordinates (x,y,z) given an inter-frame 2D point correspondence $(u,v) \leftrightarrow (u',v')$, and the camera's motion $(\mathbf{R}\,\mathbf{t})$, in the context of monocular vision.

As in a practical case, the back-projected rays cannot be expected to meet at an exact point in 3D space, an error metric has to be adopted. The triangulation procedure then looks for the best solution (x,y,z) that minimises the defined error. A reasonable choice is to find the 3D point which has the shortest Euclidean distances to both of the back-projected rays. In such a case, the error is defined as follows with respect to free parameters $k, k' \in \mathbb{R}^+$:

$$\delta_{\text{mid}}(k,k') = \|k\mathbf{a} - (k'\mathbf{a}' + \mathbf{c}')\|^2 \tag{13}$$

where $\mathbf{a} = \mathbf{K}^{-1}(u,v,1)^\top$ and $\mathbf{a}' = \mathbf{R}^\top \mathbf{K}^{-1}(u',v',1)^\top$ are the directional vectors of the back-projected rays, and $\mathbf{c}' = -\mathbf{R}^\top \mathbf{t}$ is the new camera centre (i.e. principle point) as seen in the last position's coordinate system.

The minimum of Eq. (13) can be found by calculating the least-squares solution of the following linear system:

$$\begin{bmatrix} \mathbf{a} & -\mathbf{a}' \end{bmatrix} \begin{bmatrix} k \\ k' \end{bmatrix} = \mathbf{A} \begin{bmatrix} k \\ k' \end{bmatrix} = \mathbf{c}' \tag{14}$$

The resulting values k and k' denote two points on each of the back-projected rays at the shortest mutual distance in 3D space, and the midpoint of them is therefore the optimal solution subject to the defined error metric. In particular, we have that

$$(x,y,z)^\top = \frac{1}{2}\left(\begin{bmatrix} \mathbf{a} & \mathbf{a}' \end{bmatrix} (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top + \mathbf{I} \right) \mathbf{c}' \tag{15}$$

This approach is known as *mid-point triangulation*.

If the noise of the correspondence $(u,v) \leftrightarrow (u',v')$ is believed to be Gaussian, it is proper to alternatively adopt the so-called *optimal triangulation* method. The MLE of the triangulated coordinates is achieved by minimising

$$\delta_{\text{optimal}}(\hat{\mathbf{x}}, \hat{\mathbf{x}}') = \|(u,v)^\top - \hat{\mathbf{x}}\|_\Sigma^2 + \|(u',v')^\top - \hat{\mathbf{x}}'\|_\Sigma^2 \tag{16}$$

subject to the epipolar constraint $\hat{\mathbf{x}}'^\top \mathbf{F} \hat{\mathbf{x}} = 0$, with $\hat{\mathbf{x}} = \pi_{\mathbf{K}}\left[(x,y,z)^\top\right]$ and $\hat{\mathbf{x}}' = \pi_{\mathbf{K}}\left[\mathbf{R} \cdot (x,y,z)^\top + \mathbf{t}\right]$ being the projections of the estimated 3D point.

Equation (16) poses a quadratically-constrained minimisation problem which, unfortunately, has no close-form solution. In recent years, several strategies have been developed to iteratively approach an optimal solution (see (Wu et al., 2011) for an example.)

Implementation. Based on the models described so far, monocular vision ego-motion estimation algorithms have been designed. To acquire inter-frame

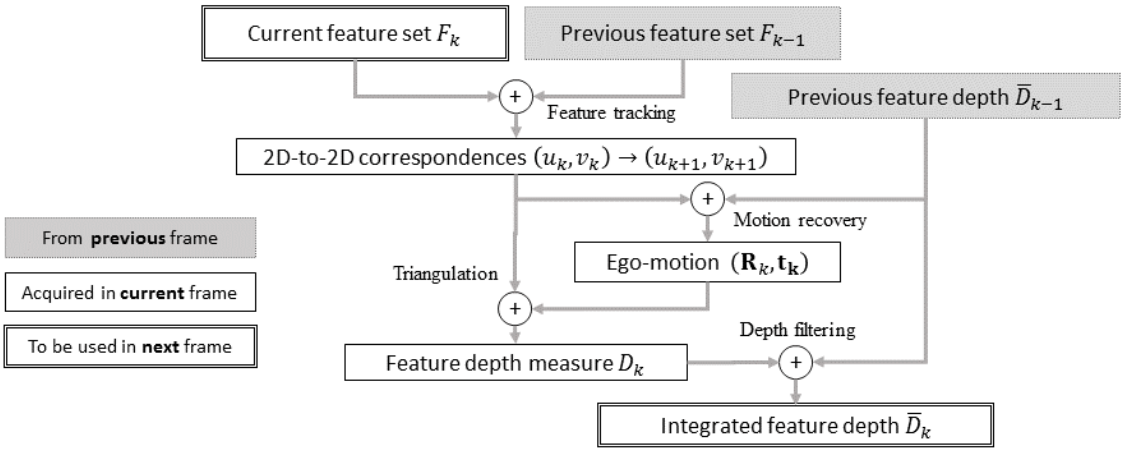

Figure 2: The pipeline of an implemented ego-motion estimator, based on the models described in Section 2.

pixel correspondences, two approaches might be considered. The first one is using all the intensities to match image blocks and produce dense correspondences; this is known as the *patch-based* technique (e.g.(Forster et al., 2014)). Alternatively, one may compare fewer but characteristic representative regions to establish sparse correspondences; this is known as the *feature-based* approach. In this section we outline an implementation based on the later technique.

In order to estimate the motion of a camera, between frame k and $k+1$, we first detect feature point sets F_k and F_{k+1}, respectively, from these two frames. The feature vectors (or feature descriptors) of these sets are then computed and matched in a high-dimensional feature space $\mathbb{R}^n$ (usually $n > 50$), by means of the Euclidean metric.

As the 3D information is not available initially, a bootstrapping technique is required to initiate the ego-motion estimation process. This can be done by applying the techniques described in Section 2 on the matched pixel correspondences in frames $k = 0$ and $k = 1$. The resulting motion $(\mathbf{R}_1\,\mathbf{t}_1)$ can then be used to triangulate the 3D coordinates of the i-th pixel correspondence $(u_{i,0}, v_{i,0}) \leftrightarrow (u_{i,1}, v_{i,1})$.

As the camera moves to the next position for frame $k = 2$, the previously triangulated 3D coordinates are used with the newly discovered pixel correspondences to recover the motion, based on the linear initialisation and non-linear minimisation models introduced in Section 2.

It is common to see a scene point involved in the ego-motion estimation in multiple frames through a sequence. This results in multiple depth estimations for the same point. Due to the error of the estimated ego-motion, the error of feature matching, and the numerical stability of the adopted triangulation method, calculated depths differ for a considered scene point, once aligned to the same coordinate system.

A depth filtering technique, in this case, may be used to fuse these measures and yield a more robust result. The recently proposed *multi-frame feature integration* (MFI) technique (Badino et al., 2013) and Kalman filter-based solutions (e.g. (Geng et al., 2015; Klette, 2014; Morales and Klette, 2013; Vaudrey et al., 2008)) are good choices.

Based on the ideas presented in this section, one may implement an ego-motion estimator which follows the pipeline illustrated by Figure 2. We leave the discussion regarding the depth integration step to the next section.

3 PROPOSED METHOD

In this section we introduce a regularised energy model to achieve more robust ego-motion recovery. An iterative depth-integration technique is also presented to further improve the performance of the motion estimation process, as more data are gathered through the sequence.

Regularised Energy Model. The idea of regularisation is to use not only 3D-to-2D point correspondences $(x_k, y_k, z_k) \leftrightarrow (u_{k+1}, v_{k+1})$ but also 2D-to-2D mappings $(u_k, v_k) \leftrightarrow (u_{k+1}, v_{k+1})$ to evaluate a motion hypothesis $(\mathbf{R}_k\,\mathbf{t}_k)$ from frame k to $k+1$.

The decision of an Euclidean transform $(\hat{\mathbf{R}}\,\hat{\mathbf{t}})$ between two views immediately instantiates an epipolar geometry, encoded by the fundamental matrix

$$\hat{\mathbf{F}} = \mathbf{K}^{-\top}[\hat{\mathbf{t}}]_{\times}\hat{\mathbf{R}}\mathbf{K}^{-1} \tag{17}$$

Intuitively one may take into account the deviation of the observed correspondences from the epipolar constraint imposed by $\hat{\mathbf{F}}$ during the energy minimisation

process. That is, in addition to the reprojection error, the minimisation now also considers the regularisation term

$$\phi_{\mathrm{E}}(\hat{\mathbf{R}},\hat{\mathbf{t}}) = \sum_i \left(\mathbf{x}'^{\top}_{i,k+1} \cdot \hat{\mathbf{F}} \cdot \mathbf{x}_{i,k}\right)^2 \tag{18}$$

where $\mathbf{x}_{i,k} = (u_{i,k}, v_{i,k}, 1)^{\top}$. Such modelling, however, is found biased and tends to move the epipole toward the image centre, as the algebraic term $\mathbf{x}'^{\top}\hat{\mathbf{F}}\mathbf{x}$ is not geometrically meaningful (Zhengyou, 1998).

A proper way is to measure the shortest distance between $\mathbf{x}'$ and the corresponding epipolar line $\mathbf{l} = \mathbf{F}\mathbf{x} = (l_0, l_1, l_2)^{\top}$ in the image plane:

$$\delta(\mathbf{x}', \mathbf{l}) = \frac{|\mathbf{x}'^{\top}\mathbf{F}\mathbf{x}|}{\sqrt{l_0^2 + l_1^2}} \tag{19}$$

The observation $\mathbf{x}'$ also introduces an epipolar constraint on $\mathbf{x}$ which yields a geometric distance

$$\delta(\mathbf{x}, \mathbf{l}') = \frac{|\mathbf{x}'^{\top}\mathbf{F}\mathbf{x}|}{\sqrt{l'^2_0 + l'^2_1}} \tag{20}$$

where $\mathbf{l}' = \mathbf{F}^{\top}\mathbf{x}'$ denotes the epipolar line in the first view. By applying symmetric measurements on the point-epipolar line distances, the energy function defined by Eq. (18) is now revised as follows:

$$\phi_{\mathrm{E}}(\hat{\mathbf{R}},\hat{\mathbf{t}}) = \sum_i \delta^2(\mathbf{x}'_{i,k+1}, \hat{\mathbf{F}}\mathbf{x}_{i,k}) + \delta^2(\mathbf{x}_{i,k}, \hat{\mathbf{F}}^{\top}\mathbf{x}_{i,k+1}) \tag{21}$$

This yields geometric errors in pixel locations.

A noise-tolerant variant is to treat the correspondence $\mathbf{x} \leftrightarrow \mathbf{x}'$ as a deviation from the ground truth $\hat{\mathbf{x}} \leftrightarrow \hat{\mathbf{x}}'$. When the differences $\|\mathbf{x} - \hat{\mathbf{x}}\|$ and $\|\mathbf{x}' - \hat{\mathbf{x}}'\|$ are believed to be small, the sum of squared mutual geometric distances can be approximated by

$$\delta^2(\hat{\mathbf{x}}, \hat{\mathbf{l}}') + \delta^2(\hat{\mathbf{x}}', \hat{\mathbf{l}}) \approx \frac{(\mathbf{x}'^{\top}\mathbf{F}\mathbf{x})^2}{l_0^2 + l_1^2 + l'^2_0 + l'^2_1} \tag{22}$$

where $\hat{\mathbf{l}} = \mathbf{F}\mathbf{x}$ and $\hat{\mathbf{l}}' = \mathbf{F}^{\top}\mathbf{x}'$ are perfect epipolar lines. This first-order approximation to the geometric error is known as the *Sampson distance* (Sampson, 1982), which has also been used to provide iterative solutions to the optimal triangulation problem as formulated by Eq. (16). When such metric is adopted in evaluating an ego-motion, Eq. (21) is formulated as:

$$\phi_{\mathrm{E}}(\hat{\mathbf{R}},\hat{\mathbf{t}}) = \sum_i \frac{(\mathbf{x}'^{\top}_{i,k+1}\hat{\mathbf{F}}\mathbf{x}_{i,k})^2}{(\hat{\mathbf{F}}\mathbf{x}_{i,k})_0^2 + (\hat{\mathbf{F}}\mathbf{x}_{i,k})_1^2 + (\hat{\mathbf{F}}^{\top}\mathbf{x}'_{i,k})_0^2 + (\hat{\mathbf{F}}^{\top}\mathbf{x}'_{i,k})_1^2} \tag{23}$$

Equations (23) and (6) are the epipolar geometry-derived energy term and the reprojection error term, respectively. By combining both equations we now model the regularised motion estimation objective function as follows:

$$\Phi(\hat{\mathbf{R}},\hat{\mathbf{t}}) = (1-\alpha)\cdot\phi_{\mathrm{R}}(\hat{\mathbf{R}},\hat{\mathbf{t}}) + \alpha\cdot\phi_{\mathrm{E}}(\hat{\mathbf{R}},\hat{\mathbf{t}}) \tag{24}$$

where a chosen damping parameter $\alpha = [0,1]$ controls the weight of the epipolar constraint.

As the 3D coordinates of a newly discovered feature are not known before the ego-motion is solved, ϕ_{R} always has fewer terms than ϕ_{E}. We therefore consider the numbers of terms in ϕ_{R} and ϕ_{E} to normalise the damping parameter. In particular, let N_{R} be the number of 3D-to-2D correspondences and N_{E} for the 2D-to-2D ones, it defines the ratio

$$\beta = \frac{N_{\mathrm{E}}}{N_{\mathrm{R}}} \cdot \frac{1-\alpha}{\alpha} \tag{25}$$

and the normalised damping parameter applied to Eq. (24) is decided by:

$$\alpha' = \frac{1}{1+\beta} \tag{26}$$

In the experiments, we investigate the effect of different α values.

Linear Initialisation and Outlier Rejection. To solve Eq. (24), the regularised energy function using an iterative least-squares minimiser, an initial guess has to be established. As an inverse problem, the ego-motion estimation problem is inherently ill-posed, and it is therefore crucial to start the optimisation with a reasonably good guess. Common initialisation strategies include linear estimation, use of a previously optimised solution and random generation. In this work we deploy a robust two-stage linear initialisation technique.

In the first stage we determine parameters $(\hat{\mathbf{R}}\ \ \hat{\mathbf{t}})$ from the essential matrix $\hat{\mathbf{E}}$, which satisfies a maximal number of epipolar constraints given by all the image-to-image observations $(u,v) \leftrightarrow (u',v')$. An observation is considered to *agree with an essential matrix* if its Sampson distance [see Eq. (22) for the definition] is within a tolerable range ε.

To avoid exhaustive search, we randomly select eight points from the observations and calculate a candidate essential matrix using the method described in Section 2. The candidate is then tested with all the observations to conclude the number of inliers. The sampling process is repeated until significantly many inliers are found within a defined limit of trials, and the best candidate is used later to initialise the optimisation process. Such a process is known as the *random sampling consensus* (RANSAC) algorithm (Fischler and Bolles, 1981).

As the translation vector $\hat{\mathbf{t}}$ obtained from the essential matrix decomposition does not provide an absolute scale, in the second stage we use 3D-to-2D correspondences $(x,y,z) \leftrightarrow (u',v')$ to recover its scale. Let k be the scale to be determined. By Eq. (1) it follows that

$$\begin{pmatrix} u' \\ v' \\ 1 \end{pmatrix} \sim \mathbf{K} \left[\hat{\mathbf{R}} \begin{pmatrix} x \\ y \\ z \end{pmatrix} + k\hat{\mathbf{t}} \right] \tag{27}$$

Let $\mathbf{a} = \mathbf{K}^{-1}(u,v,1)^\top = (a_0,a_1,a_2)^\top$, $\hat{\mathbf{R}} = (\hat{\mathbf{r}}_0,\hat{\mathbf{r}}_1,\hat{\mathbf{r}}_2)^\top$ and $\hat{\mathbf{t}} = (t_0,t_1,t_2)^\top$, Eq. (27) leads to two constraints, namely

$$(a_2t_0 - a_0t_2)\cdot k = (a_0\cdot\hat{\mathbf{r}}_2 - a_2\cdot\hat{\mathbf{r}}_0)^\top \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{28}$$

and

$$(a_2t_1 - a_1t_2)\cdot k = (a_1\cdot\hat{\mathbf{r}}_2 - a_2\cdot\hat{\mathbf{r}}_1)^\top \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{29}$$

We select a subset of the 3D-to-2D correspondences to populate an over-determined linear system of unknown k based on these constraints, and find the least-squares solution to recover the scale of $\hat{\mathbf{t}}$. The scaled ego-motion $(\hat{\mathbf{R}} \;\; k\hat{\mathbf{t}})$ is then applied to evaluate the re-projection error for each correspondence. Following the manner similar to the random sampling deployed in the previous stage, a robust estimation of k is established, with outliers identified. In the following optimisation process, all the outliers found in the initialisation stages are excluded.

Depth Integration. After introducing the term of the epipolar energy, we also like to improve the modelling of the reprojection term, which is based on 3D-to-2D correspondences. In experiments we observed that, under particular geometrical configurations, triangulated coordinates are impacted by significant nonlinear anisotropic errors. If not dealt with properly, such a depth error leads to bad ego-motion estimation. In this paper we follow a multi-frame integration strategy to temporally improve the depths of the tracked feature points.

An effective integration technique is to maintain a weighted running average of the state for each tracked feature. Let $\mathbf{m}_{i,k}$ be an observed state vector of feature i in frame k, and $\omega_{i,k} \in [0,1]$ the weight denoting how likely the observation is believed to be the true state, the estimate of the true state is calculated as

$$\bar{\mathbf{m}}_{i,k} = \frac{\bar{\omega}_{i,k-1}\cdot f_{k-1,k}(\bar{\mathbf{m}}_{i,k-1}) + \omega_{i,k}\cdot\mathbf{m}_{i,k}}{\bar{\omega}_{i,k}} \tag{30}$$

where

$$\bar{\omega}_{i,k} = \bar{\omega}_{i,k-1} + \omega_{i,k} \tag{31}$$

is the running weight and $f_{k-1,k}$ is a transition function of state from the previous frame $k-1$ to the current frame k. In this work, the state $\mathbf{m}$ are triangulated 3D coordinates, $f_{k-1,k}$ is the Euclidean transformation defined by the estimated ego-motion $(\mathbf{R}_k \;\; \mathbf{t}_k)$, and the weight is set to be $\omega_{i,k} = \frac{1}{1+\delta_{i,k}}$ where $\delta_{i,k}$ is the estimated error of the triangulation. In the case of mid-point triangulation, we use the sum of the shortest distances from a triangulated point to the two corresponding back-projected rays.

4 EXPERIMENTAL RESULTS

We report about an evaluation of the proposed model for a test sequence from the KITTI benchmark suite (Geiger et al., 2013). The sequence presents a complex street scenario, with pedestrians, bicyclists and vehicles moving in the scene. The test vehicle travelled 300 metres and captured 389 frames. We used only the left greyscale camera to calculate the ego-motion of the vehicle.

In each frame, the *speeded-up robust image features* (SURF) are detected and extracted. Features in each consecutive frame are initially matched in the feature space in a brute force manner, then outliers are identified using the RANSAC technique described in the previous section, with the tolerance distance ε set to 0.2 pixel. Before the RANSAC process begins, we augment the 2D-to-2D correspondences by performing the Kanade-Lucas-Tomasi (KLT) point tracker (Tomasi and Kanade, 1991) on the image features in frame k, which failed to find good matches in frame $k+1$. The point tracker applies a backward tracking to ensure the consistency of a correspondence, and the same tolerance distance ε is used as the threshold to reject a false match.

To evaluate estimated ego-motion in a consistent metric space, readings of the *inertial measurement unit* (IMU) of the first two frames are used to bootstrap the VO procedures. To study how the epipolar regularisation affects the accuracy, we test different values of the damping parameters $\alpha \in \{0.00, 0.25, 0.50, 0.75, 0.90\}$.

We exclude the configuration $\alpha = 1.0$ as it discards all the re-projection constraints and prohibits the ego-motion estimation in the Euclidean space. As the RANSAC technique introduces a stochastic process, for each configuration we repeated the VO procedure for 10 times and report only the best estimations in this section. Neither global optimisation (bundle adjustment) nor loop closure was used in the experiments. Two of the estimated vehicle trajectories are visualised in Fig. 3.

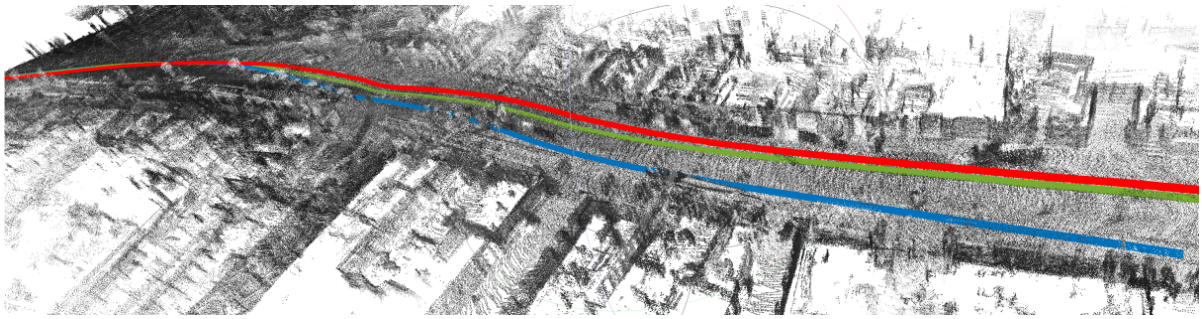

Figure 3: Visualisation of the ego-motion estimated without regularisation $\alpha = 0$ (blue) and with regularisation $\alpha = 0.9$ (green). The red line shows the ground truth motion from GPS/IMU data.

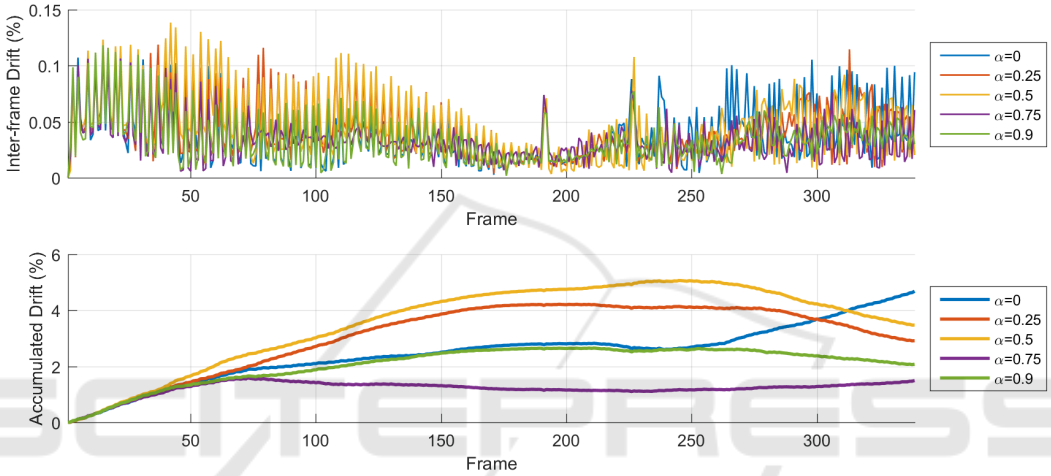

Figure 4: Inter-frame drift (*top*) and accumulated drift (*bottom*) plots of the tested sequence.

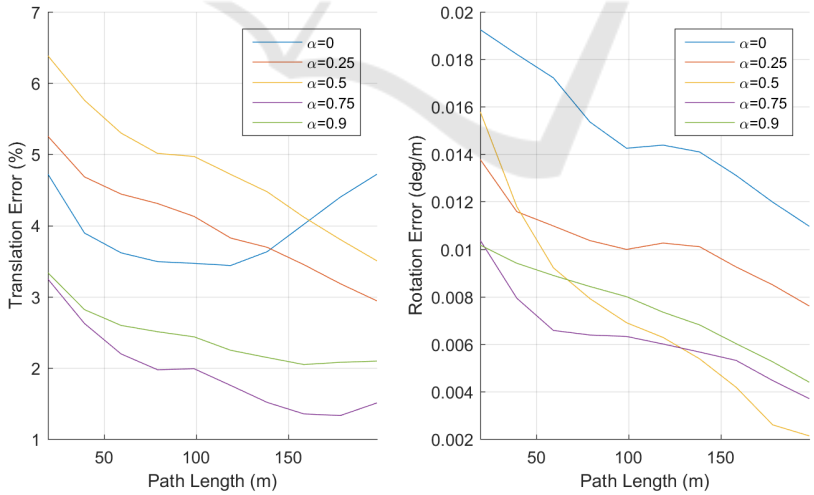

Figure 5: The errors of ego-motion of the translation part (*left*) and the rotation part (*right*) with respect to travel distance.

The drifts of the estimated vehicle position are plotted in Fig. 4. The accumulated error plot shows, with the regularisation term enabled, that the drift steadily converged to a lower bound, as observed in all the four cases where the epipolar constraints took place during the optimisation phase. We found that, as more and more pedestrians present in the field of view, the conventional approach (without regularisation) starts to deviate from the ground truth. This is shown in the inter-frame drift plot, from frame 260

thru the end of sequence. A possible reason being that, those feature points belonging to the moving objects are falsely triangulated and tracked. Arriving at the end of the sequence, the regularised energy model with α set to 0.75 achieved the lowest drift within 1.7% (3 metres), while the conventional re-projection error minimisation approach presented the worst result, with a motion drift above 5%.

We also calculated segmented motion errors in terms of translation and rotation components of the estimated ego-motion, with respect to various travel distances. The travel distance is not measured only from the beginning of the sequence; any segment begins from an arbitrary frame k thru frame $k+n$ where $n > 0$ having a length l is taken into account during the error calculation of interval $[l_p, l_q)$ if $l_p \leq l < l_q$. We divide the length of the sequence into 10 equally spaced segments for plotting. The results are depicted in Fig. 5. It shows that, in the translation component, the damping parameter $\alpha = 0.75$ yields the best accuracy in all segments, while the conventional model maintains a moderate accuracy in travel distances shorter than 100 metres. On the rotation part, however, it presents the worst accuracy. The best accuracy, achieved by $\alpha = 0.5$, which equally relies on both re-projection and epipolar constraints, is five times better than the conventional model.

5 CONCLUSIONS

In this paper we reviewed the underlying mathematical models of the monocular ego-motion estimation problem and formulated an enhanced minimisation model to improve the stability and robustness of the optimisation process.

The experimental findings support a positive effect of the proposed model on increasing the accuracy of the VO procedure. Remarkably, with monocular vision the presented implementation achieves an overall motion drift within 2% over 200 metres, which is comparative to the stereo VO implementations as listed on the website of the KITTI visual odometry benchmark in 2016.

REFERENCES

Badino, H., Yamamoto, A., Kanade, T.: Visual odometry by multi-frame feature integration. Int. ICCV Workshop Computer Vision Autonomous Driving (2013)

Engels, C., Stewenius, H., Nister, D.: Bundle adjustment rules. In Proc. Photogrammetric Computer Vision (2006)

Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Comm. of the ACM, vol. 24, no. 6, pp. 381–395 (1981)

Forster, C., Pizzoli, M., Scaramuzza, D.: SVO: Fast semi-direct monocular visual odometry. In: Proc. IEEE Int. Conf. Robotics Automation, pp. 15–22 (2014)

Geiger, A., Lenz, P., Stiller, C., and Urtasun, R.: Vision meets robotics: The KITTI dataset. Int. J. Robotics Research, vol. 32, no. 11, pp. 1231–1237 (2013)

Geng, H., Chien, H.-J., Nicolescu, R., Klette, R.: Egomotion estimation and reconstruction with Kalman filters and GPS integration. In: Proc. Computer Analysis of Images and Patterns, vol. 9256, pp. 399–410 (2015)

Hartley, R. I., Zisserman, A. : Multiple View Geometry in Computer Vision, second edition. Cambridge University Press, Cambridge (2004)

Hu, G., Huang, S., Zhao, L., Alempijevic A., Dissanayake, G.: A robust RGB-D SLAM algorithm. In Proc.: IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 1714–1719 (2012)

Klette, R.: Concise Computer Vision. Springer, London (2014)

Konolige, K., Agrawal, M.: FrameSLAM: From bundle adjustment to real-time visual mapping. IEEE Trans. Robotics, vol. 5, no. 24, pp. 1066–1077 (2008)

Lepetit, V, Moreno-Noguer, F., Fua. P.: EPnP: An accurate O(n) solution to the PnP problem. Int. J. Computer Vision, vol. 81, pp. 155–166 (2009)

Levenberg, K.A.: Method for the solution of certain non-linear problems in least squares. The Quarterly Applied Math., vol. 2, pp. 164–168 (1944)

Morales, S., Klette, R.: Kalman-filter based spatio-temporal disparity integration. Pattern Recognition Letters, vol. 34, no. 8, pp. 873–883 (2013)

Sampson, P.D.: Fitting conic sections to 'very scattered' data: An iterative refinement of the Bookstein algorithm. Computer Graphics Image Processing, vol. 18, no. 1, pp. 97–108 (1982)

Scaramuzza, D., Fraundorfer, F.: Visual odometry: Part I - The first 30 years and fundamentals. IEEE Robotics Automation Magazine, vol. 18, pp. 80–92 (2011)

Tomasi, C., Kanade, T.: Detection and tracking of point features. Carnegie Mellon University Technical Report, CMU-CS-91-132 (1991)

Vaudrey, T., Badino, H., Gehrig, S.: Integrating disparity images by incorporating disparity rate. In. Proc. Robot Vision, LNCS 4931, pp. 29–42 (2008)

Wu, F.C., Zhang, Q., Hu, Z.Y.: Efficient suboptimal solutions to the optimal triangulation. Int. J. Computer Vision, vol. 91, no. 1, pp. 77–106 (2011)

Zhengyou, Z.: Determining the epipolar geometry and its uncertainty: A review. Int. J. Computer Vision, vol. 2, no. 27, pp. 161–198 (1998)